ARCHIVES AND MANUSCRIPTS, 2017
VOL. 45, NO. 2, 77–99
https://doi.org/10.1080/01576895.2017.1330158

Routledge
Taylor & Francis Group

Binary trees? Automatically identifying the links between born-digital records

Ross Spencer

Archives New Zealand, Wellington, New Zealand

ABSTRACT

The sheer volume of records that government organisations, and thus government archives, work with on a daily basis means that there is a chance that relationships between individual records will not easily be captured and recorded. This paper begins by suggesting that the relationships described in archival catalogues will remain at the highest levels of abstraction unless they can be extracted using automated methods. Relationships that can be generated automatically are described in this paper. They will likely be less established than archivists are traditionally used to working with. For example, a so-called 'fuzzy matching' technique is discussed that may reveal the 'points' of similarity between two records. Extensible databases will be needed to store new links; flexible interfaces will be required to display them. This paper discusses some of the techniques that may currently be available for automatically identifying links between born-digital records by looking at what can be found in the data stream and the relationships digital formats inherently describe. The mechanisms described may be useful for sentencing as well as cataloguing and description. While one size will not fit all, some collections may benefit. The paper concludes by discussing briefly what this work will mean to the end user.

KEYWORDS

Item relationships; digital techniques; digital literacy; born-digital; automated description

Introduction

In an 'Internet minute', the current users of the Internet output 347,000 tweets; watch 77,000 hours of video on Netflix and communicate through 284,000 Snapchats, and these numbers are still growing.[1] Not even a fraction of that information will become part of a digital archive but it is testament to the ferocious pace at which digital records can be created through the convenience of modern computing.

We have a descriptive problem. This paper echoes current remarks such as those made by The National Archives, UK which express concern at the volume and lack of structure of born-digital record collections,[2] affecting the following functions:

- Sentencing: The application of a disposal authority and disposal actions to a class of records in an organisation.[3]

CONTACT Ross Spencer ross.spencer@dia.govt.nz

- Sensitivity review: The ability to find content in a public record that determines a restriction on who has the authority to view the record and over what period of time a restriction will be in place.
- Security: The ability to spot risks to information technology infrastructure or individuals inside a collection of public records, for example malicious content embedded in digital records.
- The ability to respond to freedom of information or official information requests, which stems from an inability to effectively extract and maintain the knowledge contained in the public record.

In the government sector, at least, there isn't the capacity to manually perform those functions, and further, provide archival description for all that is created, including connecting records to one another, without making use of a gamut of tools to do some of that work automatically.

In their wide-ranging paper Duff and Harris articulate arguments for, or against, both Fonds and Series descriptive models. Ultimately they call for more liberating methods of archival description which expose more than the archivist's voice in descriptive practices. They ask to explore 'new ways to open up archival description to other ways of representing records or naming the information in the records'.[4]

This paper partially sets itself among the arguments of Duff and Harris, and takes on board the words of Nesmith as well in suggesting that context is 'virtually boundless', and that 'more context is always needed if we are to understand what is possible to know'.[5]

By thinking about 'facts' within records that are already in the archive's control, this paper aligns itself by offering strategies which attempt to automatically extract as much information, and thus, emergent context, as possible. This information may not otherwise have been described by the archivists – nor would it be possible to utilise it so completely in the current digital era without such strategies.

The techniques can be used recursively to create connecting tissues between records from the past; records that will exist in the future; and records that exist tangentially to the content of the records we have in our possession.

These facts, or relationships, may be just one 'exploration' discussed by Duff and Harris which can 'open holes' in public records that expand users' views of them beyond generic format or subject descriptors such as 'Speeches' from a minister or an 'Annual Report' from a government agency. Relationships automatically extracted from the data available within born-digital records and recordkeeping systems can entice users into what is needed, greater discovery and discourse of archival holdings relevant to users. These facts or relationships may be just one approach to alleviating some of the concerns around the functions described by The National Archives, UK.

Archives New Zealand has demonstrated it has the capability to describe global properties of born-digital records within its adopted use of the Australian Series System; connecting records through their provenance properties such as accession number, agency, series and subseries. This paper is concerned with record-to-record relationships.

At the time of writing, item-level description at Archives New Zealand is limited to a title for the catalogue entry derived from the digital object's file name. A description consists of the time and date when the record was last modified. The content of the digital objects in

our collection remains locked away because there is not a factory floor of archivists spending their days comprehending what is inside each and every file.

Even at the beginning of the transfer process, initial, internal attempts to manually double-check the sentencing of another, external agency showed that it could take one hour to look at the content of 10 records. If it were possible to automatically identify even a handful of terms, people or places, it would be possible to garner more clues to aid this work. If it could be done for archival material, the potential would exist to link many adjacent items to each other across the breadth of the catalogue.

Back in October 2016, in a conversation with Talei Masters, Archives New Zealand's digital archivist, the following view was put forward:

> Discussions about tools and techniques for identifying relationships between items are part of a natural evolution of Archives New Zealand's descriptive thinking. In 2011 Archives New Zealand developed its new conceptual model and metadata schema for archival description. This was designed to accommodate description of born-digital records. During the drafting of the new descriptive guidelines, there was much discussion among archivists about the practicalities of describing relationships between items.
>
> Ultimately, it was acknowledged that, given the volumes of digital records likely to be in each transfer, neither agency nor Archives staff were likely to examine the content of items visually one-by-one to determine which other items they referred to. However, if this information were represented in a recordkeeping system's metadata, then it could be mapped automatically during transfer. The automated discovery of potential relationships using binary content analysis reaches to the next level; it presents an opportunity when no relationship metadata, or poor relationship metadata, exists.[6]

In this instance, Archives New Zealand identifies two potential approaches to augmenting item relationship metadata for born-digital records: where represented in external agency recordkeeping systems, capturing item relationship metadata during transfer; and binary content analysis. It is preparation for the second approach that has prompted the thinking behind this paper. This is likely to be beneficial even if the traditional data received from external agencies *is good*. The techniques here can promote the extrapolation of item-to-item relationships.

In binary is the connection between all digital objects. If digital information is looked upon as a pyramid, its foundation is a series of electronic signals interpreted as zeros and ones. Those signals can be interpreted through different encoding schemes (ASCII, EBCIDIC, Unicode), and that information is then wrapped into data structures within other data structures. The outer structure is what one might recognise as a file format – a mechanism for holding all of that information in one place for the actor (you, me, the computer) to work with. That information may be created to become a digital record.

Following this pyramid downwards, every digital record can be reduced to a series of numbers that can be re-interpreted. This has benefits because numbers are easy to compare and quantify against one other, for example less than, greater than, equal to and not equal to. In the relationship between numbers the relationship between binary objects can be discovered.

This paper will take a different approach to that of a recently proposed International Council on Archives (ICA) Records in Context (RiC) archival standard which describes 73 potential record-to-record relations,[7] and, instead of seeking an exhaustive list of every relation that might exist between two records, will seek to outline just eight relations:

(1) Relationship one: identical records.
(2) Relationship two: similar records.
(3) Relationship three: contains hyperlink.
(4) Relationship four: contains enterprise content management system (ECMS) reference.
(5) Relationship five: contains embedded digital objects.
(6) Relationship six: contains intra-item relationships.
(7) Relationship seven: contains object references.
(8) Relationship eight: item mentions.

These represent what this author believes is a pragmatic, feasible set of relationship metadata for archives to extract using free, and open-source, tools and toolchains. In the remainder of the paper the author will work through simple implementations of some of those methods.

This paper is an exploration of pragmatic approaches to extracting what we can. Different, perhaps, to the approach of the UK National Archives outlined above, which underpins research for products for eDiscovery that might be able to achieve some of the goals outlined in this paper.[8] It is a technical paper with an emphasis on open-source techniques that anyone can implement. For these reasons, it may especially be of interest to those keen on developing their levels of digital literacy.[9] It is recommended that the reader explore the contents of this paper and implement some of its ideas with a critical view to how it would operate in their own institution.

Iterative development is a creative process that builds on the continuous evaluation (successes and failures) of smaller parts of functionality. This author recommends an iterative approach to building the capability required to find relationships between records, and suggests that an organisation looking at just one relation at a time is a still a positive improvement across the recordkeeping continuum.

Throughout the paper, for the purpose of narrative variance, the terms digital object, binary object, file and digital file will be used interchangeably.

The paper begins by looking at the aforementioned binary structure of a digital object to assert identicalness between two files.

Automated methods of identifying record relationships

Checksums

A checksum is a rendition of a digital object encoded as a single, fixed-length, hexadecimal string. Checksums that are normally referred to in a digital archive also go by the name cryptographic hash function.[10] A cryptographic hash function can distil an arbitrary number of bytes down to a fixed-length hexadecimal string. Algorithms that may be adopted in the digital transfer workflow are shown in Table 1.

Table 1. Checksum algorithms and string-length comparisons.

Algorithm	String length	Author
MD5	32	Ronald Rivest, 1992
SHA1	40	US National Security Agency, 1993
SHA256	64	US National Security Agency, 2001

As an example of the differences between two algorithms, an MD5 and SHA1 checksum for the case-sensitive string *The quick brown fox jumps over the lazy dog*, looks as follows:

- MD5 – *9e107d9d372bb6826bd81d3542a419d6*
- SHA1 – *2fd4e1c67a2d28fced849ee1bb76e7391b93eb12*

We use checksums in a digital archive in order to demonstrate fixity – the quality of a record being fixed from a certain point in time, and remaining unchanged.[11] There are two features of cryptographic hash functions from their respective domains that make them particularly suited to this task:

- Unique outcome – no two distinct inputs will result in the same output.
- Deterministic outcome – every identical input generates the same output.

Given those features, it is possible to demonstrate fixity from before the point of transfer where another public sector organisation's records are taken into custody. It is also possible to demonstrate *mathematical* equivalence between two matching binary objects and show that they are identical. This grants us the opportunity to perform additional recordkeeping actions, such as 'weeding' the transfer, removing matching records (de-duplication) within the same archival context, or documenting the relationship between those two files – a relationship that says that they are identical in content but have multiple archival contexts.

To show two records have matching content, simply compare the two checksums they generate. This can be done by eye. For example, two records that produce the MD5 checksum:

900150983cd24fb0d6963f7d28e17f72

will be identical. If a single bit is changed (a file containing only the string *abc* becomes *abd*), the checksum becomes, unrecognisably:

4911e516e5aa21d327512e0c8b197616

Relationship one: identical records

As monitoring fixity via checksum is a mechanism that other archival institutions will be working with, or experimenting with already, it is the relationship of 'identicalness' that this paper wishes to highlight first.

It is not known, at the time of writing, which archives are exposing the link between identical born-digital records from their catalogue. As it is implemented, there is very little information about this being sent back and forth between Archives New Zealand's own catalogue and the digital preservation system. The latter does not contain a module for checking whether a file has been stored multiple times. The user interface of the catalogue has also not been implemented in such a way that highlights, or makes it easy to navigate, item-to-item connections.

Surfacing this information between multiple items has the potential to connect records across sub-series, and series, hierarchically in that order. It is also possible to see where identical records have been transferred by two or more separate agencies, but with good reason. For example, in an instance where agencies have previously cooperated on the same project or shared information for reference. Or it could demonstrate a breakdown in communication about recordkeeping responsibilities between the agencies.

Questions of appropriateness, as well as how to record any of the relationships shown in this paper, where to store them, and how to promote and display this information, will

be touched upon later in the paper. For now, the possibility of using these techniques will simply be registered.

Fuzzy hashing

With cryptographic hashes it is possible to demonstrate just one relationship – 'mathematical equivalence' – that is, a binary answer *yes* or *no* whether they are the same.

So-called 'fuzzy' hashing techniques exist that can compute a single fixed-length string from the binary content of a file, but with the purpose of showing the similarity of two digital files that may share a 'common ancestry' with one another.[12]

The technique comes from the field of digital forensics with the purpose of identifying deliberately changed objects by way of malice or obfuscation to avoid detection against other 'known' instances of an object.[13] An example may be a commonly used spam email 'template' where common identifiers like URL (Uniform Resource Locator: a hyperlink) or email addresses are changed. Remembering that a single bit being changed in a digital object's bitstream would unrecognisably change the output of a function such as MD5, as demonstrated earlier, a cryptographic hash cannot be used in digital forensics to spot only small changes to objects. The digital forensics use-case of spotting only the smallest changes to a digital object employs fuzzy hashing techniques to detect files designed to avoid detection. In the recordkeeping context, this technique can be utilised for more positive ends of identifying files with a common ancestry in order to enrich archival metadata.

A file that one might recognise as being 'similar' but that has small changes in its content (bitstream) may, to the reader, sound like a draft. Another example of documents with structural similarity may be the bureaucratic templates that are commonly used in an office environment, including meeting minutes, file notes and fax cover sheets.

Table 2 provides a comparison between two cryptographic hashes for a digital file picked at random.[14]

Table 2. Checksum digests are lexically different to one-another, even when differences between two files are small.

Example	Checksum
File A: with zero changes	8c69dc0668c4c73092a7042df45e756adb170742
File B: with the first byte removed	6b75b8f235c148efd1b03d9c113664895b5aa7cd

The first file is the control file. The second file has had the first byte removed. The two checksums bear no lexical relationship to one another.

In comparison, two fuzzy hashes are shown in Table 3.

Table 3. In contrast to cryptographic hashes, fuzzy hashes enables granular comparison, even when two digital files are not identical.

Example	Fuzzy hash
File A: with zero changes	1536:**t**LQy16aYRCWYTESSg3yDuBCwclnHpQ/B4kCK7ZBEY0t5vykp6CYP:**q**1a-YpYTESSgM2CwQGt9ZBB1U6hP
File C: with the first 250 bytes removed	1536:**C**LQy16aYRCWYTESSg3yDuBCwclnHpQ/B4kCK7ZBEY0t5vykp-6CYP:**B**1aYpYTESSgM2CwQGt9ZBB1U6hP

The first file is a control object, File A from Table 2, and the second has had the first 250 bytes removed. Only two characters have been changed in the output (highlighted). The algorithm used to compute these hashes – SSDEEP – outputs a score of 99 out of a possible 100, meaning that these two files are closely related to one another.[15] In the simplest terms, the similarity of the two strings produced represents how similar the two files are.

Relationship two: similar records

The authors of another algorithm for fuzzy matching, TLSH, suggest that fuzzy hashing techniques should be 'tuned for each application'.[16] Tuning requires sampling results gathered across a corpus to determine score thresholds that indicate similarity according to the group of records (or unit of content) that the technique is being used for. To provide an example, a collection of memos using the same document template may be much closer in similarity according to the fuzzy hashing algorithm, and so a narrower variance will need to be inspected to identify similarity especially when scores are high. Groups of records that are dissimilar in type will require looking at a wider variance of scores. Therefore it is important to note that the suggested use, and subsequent interpretation of, such techniques come with a prejudicial warning. Notwithstanding, the fuzzy hash can give us a second 'automatic' relationship between records, namely similarity.

As may have been implicit in the examples provided earlier, it is possible to generate hashes for information for any unit of information that the user wishes to. A checksum will normally be computed on the entirety of the file received from an agency. The same can be done for fuzzy hashing techniques. However, a fuzzy hash is not impeded by more granular changes to a file's bitstream, that is, 'the quick brown fox…' (all lower case) would theoretically produce a similar fuzzy hash to 'The quick brown fox…' (with a capitalised 'The'). There is an opportunity to make a more fine-grained comparison of born-digital records and, as a result, generate more finely tuned relationships between items. The benefit of doing so may be found when comparing, for example, a report written and transferred in Microsoft Word with one, transmitted to other users, as Portable Document Format (PDF). Comparing the information content of these two files would free the comparison from any structural bias caused by its encoding scheme.

The algorithms for generating fuzzy hashes previously discussed are both available within free, open-source tools. Content extraction tools such as Apache Tika[17] also exist that provide command line interfaces for interaction with the information encoded within a digital file. It is possible to perform some of these analyses with little or no programming ability by combining Tika to extract content with one of the tools available that implements a suitable hash algorithm, using a mechanism such as a simple batch script or shell script.[18]

Fuzzy hashes can be computed at the point a file is created, at the point the file is stored in a digital archive, or created at a time when an analyst works with it. The comparison of two fuzzy hashes for two files can also be done independently at any time, and the result could be made persistent with storage somewhere deep in the database infrastructure of the institution.

With the potential to expose further item-to-item relationships, deeper than identicalness – and with the potential to show relations across series, and across holdings – the inclusion of fuzzy hashes in our work with the born-digital record may be one worth discussing if supported by appropriate research and investigation and pending appropriate 'tuning'.

External dependencies: the HTTP:// link

The relationships discussed so far have been technical ones, mathematical distillations of a binary stream into a fixed-length string for comparison against other similar strings. The title of this paper, 'Binary Trees', alludes to other branches that can be followed to computationally identify links between born-digital records. Three links will be discussed that may be discovered in a record that can connect an item to external entities that may or may not be records under our control.

Linked items may not be under our control, and may be examples of other types of objects that an organisation might consider archiving. It may simply be something to record for making future decisions. The first such relationship this paper will talk about is discovery of Internet hyperlinks.

Burnhill, Mewissen and Wincewicz look at the challenges of hyperlinks in scholarly communication. A good indication of the scale at which hyperlinks are being used in one domain is referenced in their 2014 study. Out of 6400 e-theses, 46,000 links pointed outwards to an external reference.[19]

A rudimentary 'bash' script (a type of shell script mentioned earlier, in this case for Linux) was created to demonstrate what could be found if the content stored in Microsoft Word in the Government Digital Archive at Archives New Zealand was queried for hyperlinks:

```
#!/bin/bash

set -e

#FILES LOCATION

FILES='/home/digital-preservation/accessions'

dp_analysis ()

{

echo -e $(catdoc "$file" | grep "http://") | tr -d '[:cntrl:]'

echo

}

# Find loop...

oIFS=$IFS

IFS=$'\n'

time(find "$FILES" -type f | while read -r file; do

 dp_analysis "$file"

done)

IFS=$oIFS
```

The tool used to look inside the files here is called 'catdoc' (underlined in the example) and was capable of looking inside approximately 5297 Microsoft Word files out of 5633 files of

various file formats.[20] Combined with some manual data clean-up, approximately 4800 written lines were found to contain hyperlinks.[21]

The technique can be refined as readers start playing around with, and exploring some of, the approaches outlined in this paper. Apache Tika, mentioned earlier, was used in an attempt to do this; it was wrapped in a tool created for this paper called tikalinkextract.[22] More complicated logic was created for:

- cleaning the dataset (for example, removing inline punctuation from links);
- listing the files that links appear in;
- removing duplicates;
- finding alternative protocols such as https://.

When the same corpus of 5633 files was processed, the number returned containing hyperlinks was 942, and they contained a total of 1608 links.[23] It is possible to explain the difference through the process of cleaning links and removing duplicates, but more work will need to be done to add suitable scientific rigour to this work before reporting on it outside of the context of this paper. These two results are only intended as an indicator of work that can be done with this archival content.

Another methodology is discussed in a study by Zhou, Tobin and Grover.[24] They use the tool 'pdftohtml' with the '-xml' flag set, to convert the e-theses submitted as PDF to Extensible Markup Language.[25] This presents another method that readers can adopt to process digital content to enable the identification of external relationships. This approach also has the potential to be automated and does not require specialist technical knowledge. Zhou, Tobin and Grover discuss the challenges of using regular expressions to identify hyperlinks reliably, which is a challenge faced in the rudimentary demonstration using catdoc above; for example, line breaks in text need to be taken into account. A URL has the potential to be much more varied.[26]

Relationship three: contains hyperlink

The third relationship in this paper, a hyperlink, provides additional context to a record. A hyperlink can point to referenced webpages or other files, or records, shared on the Web. A dead hyperlink represents missing context. Exposed in the archival catalogue as a metadata entry, the hyperlink provides a search vector to other records that use the same information, and a hyperlink enables researchers to seek beyond the electronic boundaries of the catalogue search interface.

Because of the way that a hyperlink is encoded, it can be pulled out of text-based material as a relationship that can be displayed to users. Tools that support a 'Contains Hyperlink' relationship could also support better information and records management. For example:

- At sentencing it is possible to discover if there is important contextual information that is missing so that it is possible to create a strategy for replacing it, as discussed in Burnhill, Mewissen and Wincewicz.
- As the use of hyperlink citations is found to be increasing, tools can be promoted for 'permalink' generation, such as http://perma.cc suggested in Zittrain, Albert and Lessig.[27]
- For the requirements of digital preservation, the active archiving of external web links may be considered though copies may have already been recorded in the Internet Archive.[28] When context is affected by the loss of information behind the hyperlink, the web archive and government archive become intrinsically linked.

External dependencies: enterprise content management system IDs

In discussing the Web, it is useful to remind ourselves that Tim Berners-Lee's 'Modest Proposal' was for an information management system.[29] Like the Web, the modern enterprise content management system (ECMS) has been used in government offices to create links, much like hyperlinks, between and within documents. These links reduce duplication and enable better information management.

This author's understanding of an ECMS can be described simply enough as 'a formalised means of organising and storing an organisation's documents and other content, that relate to the organisation's processes'.[30] The systems one would normally expect to interact with will control metadata about that content that also includes a unique identifier for it – usually an alphanumeric token specific to the object that can externally link objects, or can be used for referencing documents inside a document's content body. The latter is something that the techniques in this paper can be used to identify.

ECMS systems that might be familiar to readers include Open Text ECMS, Lotus Document Management System and Hewlett Packard's TRIM. An example unique identifier from Objective's Enterprise Content Management system,[31] used internally at Archives New Zealand, is as follows: A123456 – consisting of a single letter 'A' followed by six digits.

Regular expressions were shown in the hyperlink example before. A regular expression describes a lexical construct that can be used to identify a known pattern inside a given text. As an example, it is possible to match a piece of text that cites an ISBN number (for example, ISBN: 978-0226143675), by using the following:

```
(ISBN:)([0-9]{3})-([0-9]{10})
```
[32]

This will identify the capitalised string 'ISBN' followed by three occurrences of the digits zero to nine, a hyphen and then ten occurrences of the digits zero to nine. It will match any string that lists a 13-digit ISBN number in this format.

A pattern can also be created to identify a well-formed ECMS reference as long as the scheme it uses is well known.

Using this technique, an Objective ECMS reference can be identified in the following phrase 'A quick brown fox saved a reference *A123456* in its ECMS' using the regular expression:

```
[[:space:]]A[0-9]{6}[[:space:]]
```

Catdoc on Linux operating systems, discussed previously, can turn a file's content into an input, via a 'pipe', for a regular expression engine called 'grep' to perform this search. This command will look as follows:

```
$catdoc <filename> | grep -E [[:space:]]A[0-9]{6}[[:space:]]
```

This matches the seven alphanumeric characters that are expected to be seen and checks that the reference is between two spaces as one might expect it to be written. Other combinations of regular expression might also work for the same string, and other approaches involving coding can be created to perform the same function with more reliability.

Relationship four: contains ECMS reference

If it is, perhaps, a tenuous link between a record sentenced for transfer to an archive and an external hyperlink, the typed, or written, relationship between two ECMS records may be more imperative to reference in archival metadata:

> Records are both evidence of business activity and information assets. They can be distinguished from other information assets by their role as evidence in the transaction of business and by their reliance on metadata.[33]

An ECMS reference for a record, inside another record's content, is an *explicitly* stated link by that record's creator, pairing the fates of those two digital objects.

While the ECMS is still active in the transferring agency, there may be a number of opportunities to make sure two records linked this way are kept together; if the time isn't taken to identify this relationship, or record it, there is a risk of losing it altogether.

Developing the capability to both identify and store the relationship between two items like this will first aid us during selection for transfer, ensuring all pertinent records are marked for transfer to archives. When extracted automatically and captured and retained in archival management systems, the ECMS identifier may consequently be helpful during archival description and discovery. The most important feature of this is to provide the best chance of ensuring that records are not orphaned from one another but even if the relationship between two ECMS items is lost, at the very least the ECMS identifier is known and signposted.

External dependencies: embedded objects

To set a reasonable scope for this paper, it has focused on document-based records. It is hoped that it has been demonstrated that there is a wealth of riches that can be discovered in records if time is taken to investigate them.

This paper will go back to its use of Tika earlier, to further demonstrate the potential to identify record-to-record relationships, and now, record-component relationships based on files embedded in other files.

Tika specialises in extracting metadata and content from a handful of formats. One of overwhelming complexity that it can handle is the Microsoft Office Family; compound objects that can be described as 'a file system inside a single file'.[34]

Through the complexity that the Microsoft Office family of formats affords the user, digital objects can be embedded inside other digital objects. A common example of this is the embedding of images in reports or presentations. The Microsoft Office format family enables embedding of any file type and the objects that might be found may also be other records held by the public sector organisation or the government archive.

To access embedded digital objects using Tika, the following command can be used:

```
java -jar tika-app-1.13.jar -z <filename>
```

If embedded objects are discovered, two potential relationships can be described.

Relationship five: embedded digital objects

One definition of an embedded object that this paper aligns itself with is:

> Record Component – Part of a Record with discrete information content that contributes to the Record's physical or intellectual completeness.[35]

The born-digital record (a Microsoft PowerPoint file) associated with Archives New Zealand catalogue Item ID R24991813 contains 71 embedded files; a plethora of images, another PowerPoint file and two other compound objects,[36] each with their own metadata. The number is only presented here as a measure of the scale of record-to-component relationships

that can be found in some digital records. It may also be an interesting use-case for digital preservation.

It is because these embedded objects can be accessed that they are interesting to us. They are not on their own, necessarily, an archival record, but they are components of the original record that an end user may find value in, either for information's sake, or for remixing and reuse as better search and discovery methods are developed.

There is no mechanism to record a component relationship in the Archives New Zealand current item model, but it is an ancillary signal that can potentially be documented alongside a record's metadata that identifies and characterises embedded objects.

Relationship six: intra-item relationships

If embedded objects cannot be recorded as component parts, then a step further would be required to identify an embedded object as an existing record held within the archive, or in another records management system elsewhere. However, that is the possibility that is presented in this paper if a handful of stratagems previously discussed are supported in various born-digital workflows.

Using checksums, and fuzzy hashing, an archive processing workflow that supports either, or indeed a handful of other mechanisms discussed in this paper, it becomes trivial to take a recursive approach to query embedded objects and send them back through the same workflows to identify how they relate to material already contained in the recordkeeping system or digital archive.

Using checksum analysis, it becomes possible to understand if a record with an existing archival reference has been embedded inside another. This could be described as an 'intra-item' relationship.

Exposing an intra-item relationship by examining embedded objects brings with it the advantages already discussed, placing the item into multiple contexts across series, agency and the remainder of the item model.

Object references

An object reference is much like an embedded object, except the file is only referred to by the primary object. It sits outside of the file otherwise.

An HTML webpage will use something called an 'src' attribute within other markup tags, such as those for images (<img>), to identify the location of a given piece of media to display or play.[37] The source attribute will point to a file that is external to the HTML document itself.

As well as 71 embedded digital objects in the Microsoft PowerPoint presentation discussed in the previous example, it also contains instances of references to external objects – in this case seven videos. Those videos are available to the user when the record is viewed in the Archives New Zealand catalogue.[38]

Out of the information this paper has sought to extract from digital files so far, finding external object references may be the hardest piece of information to extract without suitable tools, but it is still feasible to do so.

Working with the example record, the video links were first observed by looking at the file in Microsoft PowerPoint. The videos which played through the file could also be seen as standalone objects in the same folder. This was enough in our ingest workflow to be able

to package the files together for submission into the digital repository. This was a manual process and the relationships could easily have gone unnoticed, How can techniques be developed to do this more reliably? How can this process be automated?

Any digital file can be viewed through the lens of a hex editor, that is, a viewer that represents binary numbers using the hexadecimal number system. To demonstrate what hexadecimal looks like, the numbers zero to sixteen can be written as follows:

00, 01, 02, 03, 04, 05, 06, 07, 08, 09, 0A, 0B, 0C, 0D, 0E, 0F, 10

The hexadecimal number system makes binary easy to read because it can represent eight binary 'bits' as a single hexadecimal value and thus enables the display of the entire underlying structure of a file.

If the Microsoft PowerPoint record above is viewed through this lens, variably sized blocks of data can be found that reference each of these videos. These blocks look like:

```
00 0F 00 05 10 34 00 00 00 00 04 10 08 00 00 00 04
00 00 00 00 00 33 01 00 00 BA 0F 1C 00 00 00 6D 00 2D
00 74 00 6F 00 76 00 2D 00 73 00 68 00 65 00 6D 00 2E
00 77 00 6D 00 76 00 0F 00 07 10 38
```

Many of these hexadecimal values will be decoded for specific purposes by the computer (represented with dots below), but other values are mapped to the Latin alphabet (also seen below in bold text):

.....4.....3.....**m.t.o.v.s.h.e.m..w.m.v**.....8

The filename of the embedded object can be seen between each of the dots as:

m-tov-shem.wmv

This particular object reference is described in the specification for Microsoft PowerPoint as ExVideoContainer.[39] Whenever a PowerPoint slide is reached containing such a reference, the file must be available in the same folder as the PowerPoint presentation so that it can be replayed to an audience.

There is still some work to be done here but, for any file that can be analysed this way, a tool can be created that answers some of the questions raised about automation above.

Relationship seven: contains object references

Relationship seven should be a familiar relationship to the reader. It exists in Dublin Core as 'hasPart'.[40] The bibliographic standard MARC also contains a hasPart-like element.[41] It has been proposed for the ICA conceptual model as a handful of record-relations and record-component relations, for example RiC-R14 'has part'.[42]

Many file formats will have objects where this relationship is implied, including popular file formats such as HTML, Scalable Vector Graphics and Microsoft PowerPoint.

The purpose of the relationship's inclusion in this paper is to make a call-to-arms for development of the capability to extract metadata about it automatically in more file formats in archival and digital preservation workflows, as in the given example.

The relationship itself has importance once again across the information and records management continuum, from sentencing to preservation within a digital archive.

Some reasons for the importance of this relationship are as follows:

- Providing the fullest context for the record, for example it becomes possible to prevent loss of this context by maintaining the record and its constituent parts.
- The inverse of this relationship means that the individual parts do not become orphaned.
- This in turn has an implication for digital preservation whereby it becomes an imperative to maintain each external link directly associated with a record, for example when migrating objects to a new representation format.

Content-level analysis

One method of accessing the content in Microsoft Word using 'catdoc' has been demonstrated in this paper. It is possible to use Apache Tika for a wider variety of formats using the '-t', or '--text' flag to output plain text.

Once information has been extracted as plain text it is possible to analyse it in a number of different ways.

This paper will briefly explore, firstly, the comparison of content to a dictionary of values, and secondly, named entity extraction. The mechanisms vary in complexity, but the relationships that can be documented are equivalent.

Content-level analysis: dictionary based

Though rudimentary in approach, it is possible to process every word in a piece of content and compare it to words and phrases available in a pre-defined dictionary of values important to the organisation, for example prime ministers and Māori organisation names. A dictionary may be a list of any key terms important to an institution's context.

Matched values can be linked back to key parts of a catalogue such as 'Agency', or would simply act as keywords, enabling faceted discovery for records mentioning the same key terms.

The simplest way to implement this search is to take the regular expression-based examples discussed previously and change the terms from patterns matching 'HTTP://' or 'A{123456}' to each subsequent term in a list: 'Julia Gillard', then 'Kevin Rudd', for example.

In the archival catalogue, or alternative discovery mechanism, the user is immediately made aware of a phrase, organisation or public figure that might be important, as a flag. An index that links out to other records mentioning the same phrase means that the entire collection becomes connected.

As the reader experiments with these techniques, they will quickly understand some of the initial limitations, such as taking into account minor spelling mistakes for terms as well as synonyms and acronyms which describe the same concepts. They will consider approaches to correct these issues and develop strategies such as fuzzy analysis, above, to counter them. In the spirit of exploration and the development of capability through iterative processes discussed in the introduction, these techniques are introduced by way of starting somewhere.

Content-level analysis: named entity extraction

Named entity extraction uses natural language processing techniques to draw information out of text, which includes, among other things, the names of people, place names, dates and organisations.

Natural language processing asks computers to interpret data as natural language, like humans do, making inferences about sentence structure which reveal nouns, verbs and so on, affording us the opportunity to output and then classify the type of information that is found.

Take an extract in Figure 1 from Wikipedia, processed through Stanford's Named Entity Recognizer tool.[43]

It is possible to find over a dozen entities to work with in this text alone, and given appropriate quality assurance checks it is possible to talk about how to incorporate this information automatically into the archival description.

This could be the intellectual content of an archival record; a minister's speech, for example. Within an Australian or New Zealand catalogue a higher precedence could be placed on terms specific to that location, and could be treated as such. Though it might not make sense to model entities for New York City, or the Metropolitan Life Insurance Company Tower as in the example below, it represents a potential to add many dimensions to archives' holdings. The records that are held by archival organisations could become multidimensional data that provides new and untrodden routes of access into collections, or indeed linked out to federated catalogues.

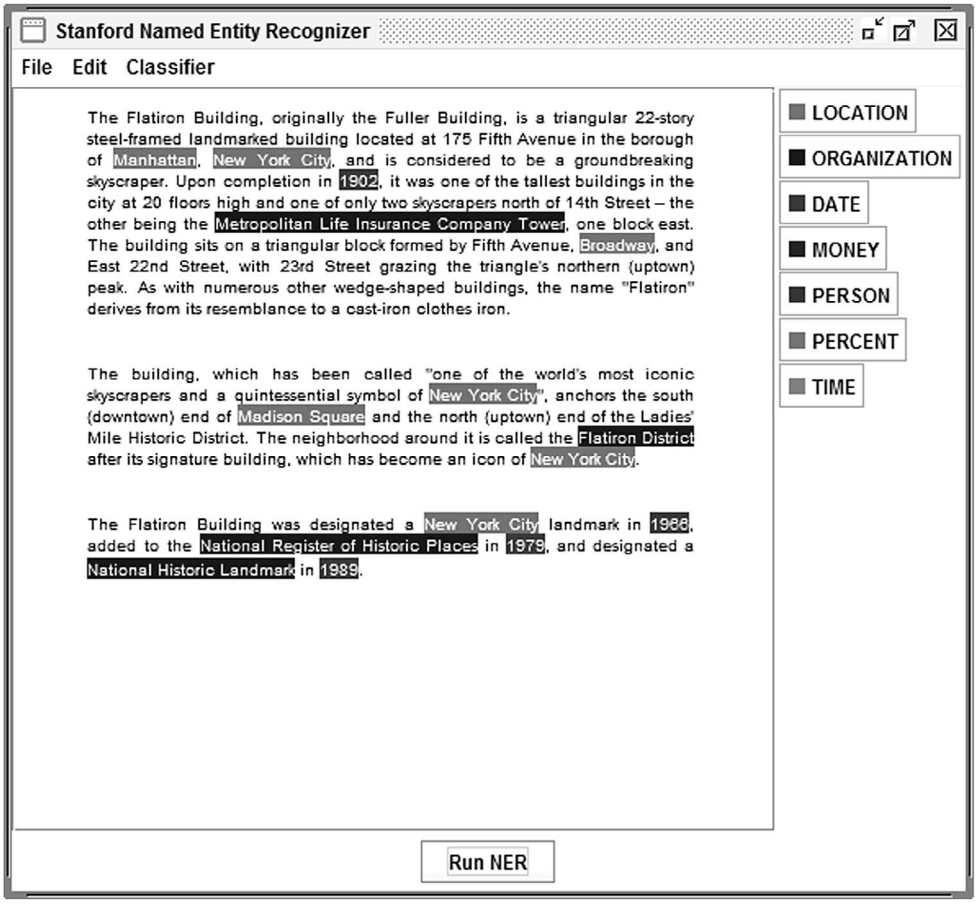

Figure 1. Example entity highlighting using Stanford's Named Entity Recognizer.

Relationship eight: item mentions

It is likely that there will be a temptation to wrap greater archival context around entities extracted for an 'Item Mentions' relationship, but this should not be a barrier to surfacing this data. The suggestions of this paper are pragmatic. Provide access to the data for which there is a low barrier to access. Use low-cost tools such as catdoc, Tika, antiword,[44] grep and shell-scripting. Process that data with other free, and open-source tools providing named-entity extraction, Python's Natural Language Toolkit,[45] or indeed, Tika again.[46] Use the output and light quality assurance to provide a bit more information.

Key-term indexing may lack sophistication in comparison to other methods of archival description, but following in the footsteps of the tag-cloud in modern blogs, and hashtags in tweets, it simply provides an additional method for knowledge organisation;[47] a platform to annotate records with more detailed description; and an additional way for users to reach into catalogue-based collections.

Discussion

The introduction to this text referred to Duff and Harris's discussion of the importance of liberatory methods of description within the archive to be able to hear more than the archivist's voice and to find new ways to open up archival description.[48] This text has outlined a set of methods that can begin a conversation about how practically to achieve this.

For one reason above all, because of the increasing numbers of digital records that our organisations are likely to receive, this author does not think there is a choice but to take a computational approach, including in other parts of the recordkeeping continuum.

The ICA Records in Context Initial Draft Standard provided this paper with a metric for potential item-to-item relationships, describing 73.

If context is indeed virtually boundless, as described by Nesmith in the introduction, then some of these relationships and also those in this paper will go some way to adding new contexts at the item level – between records, across series, and potentially across catalogues and collections. Record relationships have the potential to connect information but they're also difficult to describe in any great number by hand.

As an act of interpretation, a human being describing each digital record is not just cognitively exhausting, but it is also time-consuming. Take a seemingly innocuous piece of information. How would one ascribe the property of 'draft' to a record – take, for example, a minister's speech?

- Is a draft still a draft if 80% of the content is changed from the version delivered?
- Is a draft a draft because it is marked as such inline or in metadata?
- Is a draft still recognised as a draft if a record was used 'in the wild' to address the public?

It may not even be feasible to describe records at the item level by hand but at the very least it is demonstrably feasible to develop rudimentary automated approaches to help. This paper outlines eight such approaches. It is possible to give cataloguers a head start on description by using potentially objective, automated abstractions of a record, hidden within its binary stream.

This author also approaches the described methods as an act of exploration of the information held within a digital archive and suggests, almost simplistically, that it is because it

is there, and that because it is possible to extract these relationships from a record, archival institutions should.

Challenges to the various approaches

There are several challenges to the use of the methods outlined here. The first is the objectivity of any single approach and the identification and representation of human and machine biases in the application of those methods. Another is the digital literacy of the individual accessing and reading the metadata of a record. A third challenge is that database-like technologies and infrastructure must also support flexible, fluid,[49] growing archival description.

Some of the techniques discussed within this paper are entirely objective. To prove one record is identical to another in a different context, a mathematical technique is used to generate supposedly identical checksums. If, as is highlighted in the introduction to this paper, we are to expose more than the archivist's voice in descriptive practices, then it is hoped that techniques which look, specifically, at a record's content will help to do this.

It is not possible to make techniques such as those used for natural language processing entirely free from the bias of human interpretation but it may be possible to reduce it. By exploring techniques such as those used in relationship eight in this paper, 'item mentions', the likely existent bias brought about through selection and appraisal can be exposed, but also played around with; to say 'this' paper is not just a 'Minister's Annual Report' but that it is also a document from government that discusses 'this' person, or 'this' place name or 'this' object, potentially includes previously unregistered characteristics.

As this paper has emphasised, by building capability and using open-source techniques, it is hoped that as readers of this paper become more familiar and more capable with what is proposed, they can review algorithms and positively affect bias by adding to datasets, for example by adding to dictionary-based approaches to identifying entities. Or by learning how to review code-based approaches, becoming more informed about a code's flaws and coming up with strategies to influence or correct those.

An infrastructure is required that can grow as description matures and is added to. Verhoeven asks how our technical infrastructures can be reconfigured and reconfigure the sense of, and possibility for, acts of connection and the felt experience of connectedness in archives, and how 'connections–links–relations' are materialised in digital environments.[50]

This author aligns himself with the political importance of her paper and the connections that Verhoeven discusses that have the potential to 'subvert the idea of authoritative "sources" altogether'. Her proposal also relies on a flexible infrastructure, techniques that pull us away from the confined world of schemas and relational database management systems, to the world of ontologies. Specifically, vernacular ontologies that can incorporate human experience and knowledge in what is presented by an archival collection back to the world.

Ethical implications

This paper has not yet covered any of the ethical implications of working with records in any of the ways discussed. Again, it must be recognised that the techniques can be used in many different contexts: sentencing, for example (a task not normally visible to the public), or aiding in archival description, which needs to happen whether the records are open or restricted access. The access status of public records and their metadata is expected

to be maintained through jurisdictional policies. This may mean that special attention is warranted to monitor the extent of metadata made available, using some of the techniques outlined in this paper, when records are expected to be restricted.

For open-access records, then, the author's own view is that the archival institution has an opportunity to enrich access for its users. In many cases, this is likely to provide some shortcuts to results likely to be output by the digitally literate researcher capable of mixing and remixing digital records using their own tools and techniques. It is hoped that these shortcuts benefit those researchers who can push new boundaries even further. The links and connections that can be made as a result of applying the methods in this paper are merely using information already available in a record or set of public records.

Tim Sherratt notes that people are already using 'our digital stuff in ways we don't expect' and asks 'whether libraries, archives, and museums see this hunger for connection as an invitation or a threat'.[51]

If we take the position that this is an invitation, then Sherratt delves deeper into institutional concerns about opening up collections, including the desire to avoid misrepresentation, mislabelling or misuse of cultural objects. He references Dan Cohen,[52] executive director of the Digital Public Library of America, who suggests that:

> The cynics, of course, will say that bad actors will do bad things with all that open data. But here's the thing about the open web: bad actors will do bad things, regardless. They will ignore whatever license you have asserted, or use technical means to circumvent your technical lock.
>
> The flip side of worries about bad actors is that we underestimate the number of good actors doing the right thing.

In this author's opinion, we can be the first and most important 'good actor' by opening up as much content as we can to aid users in finding what they need. The most alluring benefits, echoing Duff and Harris, are further emphasised in some of Sherratt's closing remarks: 'The people buried inside a recordkeeping system can be brought at last to the surface.'

Display and discovery

As noted numerous times, the techniques discussed in this paper are applicable across the recordkeeping continuum. Of interest to this author is where techniques are required in a collecting archive, to sentence better, to describe records better and to provide better finding aids to end users. Display of information for any of these purposes is something that still needs to be explored but is largely beyond the scope of this document.

Gollins and Bayne describe abstractions of information that have become part of The National Archives, UK catalogue and finding tool, Discovery.[53] The National Archives added two such abstractions that may satisfy the end-user experience for the type of data one might expose through utilising the techniques in this paper, namely 'Information Asset View' and 'Annotations'. The former 'allows a specific subset of the properties of an asset (or indeed its parent asset) to be exposed for a particular purpose (for example … for display to a user)'. The latter 'allow[s] additional values to be added to the catalogue'.[54] Gollins and Bayne are describing the documenting of 'facts' about a record.

Both of these techniques notionally align with what might be exposed through the techniques this paper has covered. This could be most interesting to the relationship 'item

mentions', where the content is analysed for references to names, places or events, among other pieces of information.

Notably, with the exception of acknowledging graph technologies below, this paper will refrain from suggesting any particular technology, encoding, storage or interlinking method for any one piece of data. It is this author's opinion that there is no single scheme or answer that is required to make these technologies work. The author wishes to emphasise the importance of trying these techniques, looking for these kinds of information, and then taking the time to understand where they might work for your organisation. Encodings exist – but the data which the user or organisation wants to encode also has to exist. In the spirit of digital literacy, it is important to understand what might become feasible.

Building towards a solution

An iterative approach is important in this work. It is an approach that begins by gaining the organisation's support to build this capability and taking small steps towards it. The following repeating steps are needed:

- A simple mechanism to extract records from the archive for processing.
- An extensible data model that can record the new connections and relationships among those items extracted.
- Exploration of the content extracted to discover new layers of information and ideas about what else can be discovered in our collections.
- Development of policy that permits or restricts the uses of newly discovered links and connections as appropriate to a jurisdiction.
- Development of techniques to surface new information to end users, wherever they may be in the recordkeeping continuum.

Outside of the subversion of authoritative sources of information, it is the implication of the technical infrastructure that Verhoeven discusses that this paper will pay closer attention to at this point, because it is that flexibility which is almost certainly part of an iterative process required to develop the capability to present the links she discusses, and those in this paper here.[55]

Part of the infrastructure discussed by Verhoeven is in the database technology used, relaying a contemporary renaissance in database management system design, and there is also a question raised about what digital research and search functions will need to look like.

Seeing the world as connected in more different ways than can perhaps be allowed by the restrictive schemas of a relational database management system is echoed in the 2015 Australian/New Zealand recordkeeping standard AS/NZS 5478:2015, Recordkeeping metadata property reference set, which states: 'The digital world is increasingly using networked relationships.'

Although still a proposal at time of writing, the technology required to provide enough extensibility to encode such relationships is extolled in the ICA's proposed conceptual model in regard to graph technologies:

> But, given that the real world within which we live and work may be understood as a vast, dynamically interrelated network of people and objects situated in space and time, graph technologies offer new and more expressive forms of representation.

> It is within the context of established and emerging communication technologies and of an expanded understanding of provenance, that RiC-CM is being developed. RiC-CM is intended to accommodate existing description practices and at the same time to acknowledge new understandings, and to position archives to take advantage of opportunities presented by new and emerging communication technologies. RiC-CM aspires to reflect both facets of the Principle of Provenance, as these have traditionally been understood and practiced, and at the same time recognize a more expansive and dynamic understanding of provenance. It is this more expansive understanding that is embodied in the word 'Contexts.' RiC-CM is intended to enable a fuller, if forever incomplete, description of the contexts in which records emerge and exist, so as to enable multiple perspectives and multiple avenues of access.[56]

This paper suggests that work towards a multidimensional perspective on describing archives is started by taking a simple and pragmatic approach, using inexpensive tools as demonstrated, and focusing on good-quality assurance to make sure that the output first is fit for purpose before repeating and improving.

This paper purposely limits the number of relations it talks about – eight. Starting even more modestly with the born-digital information at Archives New Zealand; if the data is augmented with just one of these new relationships, it will increase the amount of information available to users to access by a factor of approximately five and a half thousand.[57]

This starts to add up, taking the approach of one relation at a time.

Our digital records are network parts with each relation forming an edge between vertices in that network. Each vertex is more than a record in its own right. It is a connection between people, places and events, and records in different dimensions reach out to other nodes providing different archival context.

Conclusion

The application of technology to describe relationships between born-digital records enables gaps to be filled where manual techniques would otherwise take too long or not be possible. In Kornblum it is stated:

> Computer forensic examiners are often overwhelmed with data. Modern hard drives contain more information that cannot be manually examined in a reasonable time period creating a need for data reduction techniques.[58]

It is the assertion of this author that each of the techniques here, at the very least, are data-reduction techniques that can be of benefit to the roles of the information and records manager, the archivist, the librarian and the digital preservation researcher.

What is more, each of these techniques is presented with low technical and cost barriers to trying out and can be found within easy reach of the institutions in which the roles of developers, archivists and researchers are supported.

The theme of this piece was how it is possible to find new links and relations between records and information. This information sits across the continuum of information and records management and as such can have a positive impact on all the functions enacted on a record across that continuum. Functions identified include sentencing, archival description, discovery by the end user and digital preservation.

For the latter function, this paper might serve as a reminder that digital preservation is not simply a discussion about looking after a bitstream, or whether or not a record can be emulated in an environment or migrated to another file format. Digital preservation is about

the preservation of vast digital networks of information that we may fall foul of forgetting if we are not actively keeping track of the relationships between every record that we look after.

It is therefore a reminder that digital preservation is information and records management, and that information and records management is also digital preservation.

Endnotes

1. Domosphere, 'Data Never Sleeps 3.0', 2015, available at <*https://www.domo.com/blog/2015/08/data-never-sleeps-3-0/*>, accessed 15 November 2016.
2. The National Archives, UK, 'The Application of Technology-assisted Review to Born-Digital Records Transfer, Inquiries and Beyond: Research Report', available at <*http://www.nationalarchives.gov.uk/documents/technology-assisted-review-to-born-digital-records-transfer.pdf*>, accessed 17 January 2016.
3. Archives New Zealand, 'Disposal–Sentencing, October 2016', available at <*http://records.archives.govt.nz/assets/Guidance-new-standard/Disposal-Sentencing-16-G10.pdf*>, accessed 11 April 2017.
4. WM Duff and V Harris, 'Stories and Names: Archival Description as Narrating Records and Constructing Meanings', *Archival Science*, vol. 2, no. 3, September 2002, pp. 263–85.
5. T Nesmith, 'Reopening Archives: Bringing New Contextualities into Archival Theory and Practice', *Archivaria*, vol. 60, Fall 2005, pp. 259–74.
6. Personal communication with Talei Masters, October 2016.
7. International Council on Archives (ICA) Experts Group on Archival Description, 'Records in Contexts – A Conceptual Model for Archival Description, Consultation Draft v0.1', 2016, available at <*http://www.ica.org/sites/default/files/RiC-CM-0.1.pdf*>, accessed 15 November 2016.
8. EDiscovery is defined as: 'The discovery or disclosure of electronic information for the purposes of litigation. This phrase is used in the United States but is also the common descriptor for software tools that assist with eDiscovery–eDisclosure in the United Kingdom', The National Archives, UK.
9. L Masterman, 'From Digital Literacy to Digital Capabilities', available at <*https://blogs.it.ox.ac.uk/acit-news/2016/05/18/dig-lit-and-dig-cap/*>, accessed 17 January 2017.
10. Wikipedia.org, 'Cryptographic Hash Function', available at <*https://en.wikipedia.org/w/index.php?title=Cryptographic_hash_function&oldid=749155568*>, accessed 15 November 2016.
11. Society of American Archivists, 'Glossary', available at <*http://www2.archivists.org/glossary/terms/f/fixity*>, accessed 15 November 2016.
12. J Kornblum, 'Identifying Almost Identical Files Using Context Triggered Piecewise Hashing', 2006, available at <*https://www.dfrws.org/sites/default/files/session-files/pres-identifying_almost_identical_files_using_context_triggered_piecewise_hashing.pdf*>, accessed 15 November 2016.
13. J Oliver, C Cheng and Y Chen, 'TLSH – A Locality Sensitive Hash', 2014, available at <*https://github.com/trendmicro/tlsh/blob/master/TLSH_CTC_final.pdf*>, accessed 15 November 2016.
14. Wikimedia Commons, 'File: A Corridor of Files at The National Archives UK.jpg', available at <*https://commons.wikimedia.org/w/index.php?title=File:A_corridor_of_files_at_The_National_Archives_UK.jpg&oldid=213314497&uselang=en-gb*>, accessed 15 November 2016.
15. J Kornblum, 'ssdeep – Latest Version 2.13', available at <*http://ssdeep.sourceforge.net/*>, accessed 15 November 2016.
16. J Oliver, S Forman and C Cheng, 'Using Randomization to Attack Similarity Digests', 2015, available at <*https://github.com/trendmicro/tlsh/blob/master/Attacking_LSH_and_Sim_Dig.pdf*>, accessed 15 November 2016.
17. Apache.org, 'Apache Tika', available at <*https://tika.apache.org/*>, accessed 15 November 2016.
18. A batch, or shell script, is an automation script, its type specific to the operating system, that literally 'scripts', in order, activities for the system to perform, examples of which are used later in the paper.

19. P Burnhill, M Mewissen and R Wincewicz, 'Reference Rot in Scholarly Statement: Threat and Remedy', 2015, available at <*http://hiberlink.org/Insight.htm*>, accessed 15 November 2016.
20. P Warden, 'catdoc', available at <*https://github.com/petewarden/catdoc*>, accessed 15 November 2016.
21. R Spencer, 'ASA Binary Trees: E-accession Hyperlinks Rudimentary Extract', 2016, available at <*https://gist.github.com/ross-spencer/a6411a021afb7de7e3dc6dd713f7b520/aa3f40dd48def93ad900e4d025ab15ab11da044d*>, accessed 9 May 2017.
22. R Spencer, 'tikalinkextract', available at <*https://github.com/httpreserve/tikalinkextract*>, accessed 9 May 2017.
23. R Spencer, 'httpreserve/eaccession-research: eAccessions Hyperlinks Version 1.0.0', 2017, available at <*http://doi.org/10.5281/zenodo.495809*>, accessed 9 May 2017.
24. K Zhou, R Tobin and C Grover, 'Extraction and Analysis of Referenced Web Links in Large-scale Scholarly Articles', 2014, available at <*http://homepages.inf.ed.ac.uk/kzhou2/papers/dl2014-zhou.pdf*>, accessed 15 November 2016. Burnhill, Mewissen and Wincewicz.
25. D Noonberg, 'PDFTOHTML', available at <*http://pdftohtml.sourceforge.net/*>, accessed 15 November 2016.
26. J Goyvaerts, 'Detecting URLs in a Block of Text', 2008, available at <*http://www.regexguru.com/2008/11/detecting-urls-in-a-block-of-text/*>, accessed 15 November 2016.
27. J Zittrain, K Albert and L Lessig, 'Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations', 2014, available at <*http://harvardlawreview.org/2014/03/perma-scoping-and-addressing-the-problem-of-link-and-reference-rot-in-legal-citations/*>, accessed 15 November 2016.
28. Internet Archive, available at <*http://archive.org*>, accessed 15 November 2016.
29. T Berners-Lee, 'Information Management: A Proposal', 1990, available at <*https://www.w3.org/History/1989/proposal.html*>, accessed 15 November 2016.
30. Wikipedia.org, 'Enterprise Content Management (ECM)', 2017, available at <*https://en.wikipedia.org/w/index.php?title=Enterprise_content_management&oldid=759565057*>, accessed 18 January 2017.
31. Objective.com, 'Enterprise Content Management', available at <*http://www.objective.com/products/enterprise-content-management*>, accessed 15 November 2016.
32. Regex101.com, 'Untitled', 2017, available at <*https://regex101.com/r/ry8aTb/1*>, accessed 19 January 2017.
33. International Standards Organisation, 'Information and Documentation – Records Management – Part 1: Concepts and Principles', 2016, available at <*https://www.iso.org/obp/ui/-iso:std:iso:15489:-1:ed-2:v1:en*>, accessed 15 November 2016.
34. J Spolsky, 'Why Are the Microsoft Office File Formats So Complicated?', 2008, available at <*http://www.joelonsoftware.com/items/2008/02/19.html*>, accessed 15 November 2016.
35. ICA Experts Group on Archival Description, p. 13.
36. Archives New Zealand, 'Digital Future Summit-video', 2007, available at <*https://www.archway.archives.govt.nz/ViewFullItem.do?code=24991813&digital=yes*>, accessed 15 November 2016.
37. W3Schools, 'HTML <img> src Attribute', 2017, available at <*http://www.w3schools.com/tags/att_source_src.asp*>, accessed 18 January 2017.
38. Archives New Zealand, 'Digital Future Summit-video – Direct Download Link', 2007, available at <*http://ndhadeliver.natlib.govt.nz/delivery/DeliveryManagerServlet?dps_pid=IE25298510*>, accessed 15 November 2016.
39. Microsoft Development Network, '[MS-PPT]: PowerPoint (.ppt) Binary File Format', available at <*https://msdn.microsoft.com/en-us/library/office/cc313106(v=office.12).aspx*>, accessed 15 November 2016.
40. Dublin Core Metadata Initiative, 'DCMI Metadata Terms: hasPart', 2012, available at <*http://dublincore.org/documents/dcmi-terms/-terms-hasPart*>, accessed 15 November 2016.
41. Development and MARC Standards Office: Library of Congress, 'MARC to Dublin Core Crosswalk', 2008, available at <*https://www.loc.gov/marc/marc2dc.html*>, accessed 18 January 2017.
42. ICA Experts Group on Archival Description, p. 40.

43. Wikipedia.org, 'Flatiron Building', 2016, available at <*https://en.wikipedia.org/w/index.php?title=Flatiron_Building&oldid=742046805*>, accessed 15 November 2016; The Stanford Natural Language Processing Group, 'Stanford Named Entity Recognizer (NER)', available at <*http://nlp.stanford.edu/software/CRF-NER.shtml*>, accessed 15 November 2016.
44. A Van Os, 'Antiword: A Free MS Word Document Reader', available at <*http://www.winfield.demon.nl/*>, accessed 15 November 2016.
45. NLTK Project, 'Natural Language Toolkit', available at <*http://www.nltk.org/*>, accessed 15 November 2016.
46. Tika Wiki, 'Named Entity Recognition (NER) with Tika', available at <*https://wiki.apache.org/tika/TikaAndNER*>, accessed 15 November 2016.
47. Twitter Help Center, 'Using Hashtags on Twitter', available at <*https://support.twitter.com/articles/49309*>, accessed 15 November 2016.
48. Duff and Harris.
49. The National Archives, 'Digital Strategy 2017', available at <*http://www.nationalarchives.gov.uk/documents/the-national-archives-digital-strategy-2017-19.pdf*>, accessed 11 April 2017.
50. D Verhoeven, 'As Luck Would Have It: Serendipity and Solace in Digital Research Infrastructure', 2016, available at <*http://www.academia.edu/21802414/As_Luck_Would_Have_It_Serendipity_and_Solace_in_Digital_Research_Infrastructure*>, accessed 15 November 2016.
51. T Sherratt, 'Life on the Outside: Collections, Contexts and the Wild, Wild Web', 2014, available at <*https://medium.com/@wragge/life-on-the-outside-collections-contexts-and-the-wild-wild-web-4d334ccddee2#.fvpib06h0*>, accessed 18 January 2017.
52. D Cohen, 'CC0 (+BY)', 2013, available at <*http://www.dancohen.org/2013/11/26/cc0-by/*>, accessed 18 January 2017.
53. T Gollins and E Bayne, 'Finding Archived Records in a Digital Age', in M Moss, B Endicott-Popovsky and MJ Dupuis (eds), *Is Digital Different?* Facet Publishing, London, 2015, pp. 128–48.
54. ibid., p. 145.
55. Verhoeven.
56. ICA Experts Group on Archival Description, p. 9.
57. Approximate number of born-digital items in the collection at the time of writing.
58. Kornblum, 'Identifying Almost Identical Files'.

Disclosure statement

No potential conflict of interest was reported by the author.

Notes on contributor

Ross Spencer has a BSc in Software Engineering and an MA in Digital Culture and Technology. At The National Archives, UK he worked in the Digital Preservation Department as a digital preservation researcher, developing the capabilities of the DROID file format identification tool. At Archives New Zealand he works as a digital preservation analyst, developing digital preservation policy as well as helping to mature the organisation's digital preservation capabilities within the System Strategy and Standards team.